DOCUMENT RESUME

ED 251 053 HE 017 919

AUTHOR Slawski, Carl
TITLE Learning versus Rating Teaching.
PUB DATE 29 Oct 84
NOTE 71p.; Paper presented at the Conference on Postsecondary Education sponsored by the Association for the Study of Higher Education and the American Educational Research Association Division J (San Francisco, CA, October 28-30, 1984).
PUB TYPE Reference Materials - Directories/Catalogs (132) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *College Faculty; Employment Practices; *Evaluation Methods; *Faculty Promotion; Higher Education; Legal Responsibility; *Personnel Policy; *Predictor Variables; *Student Evaluation of Teacher Performance; Validity

ABSTRACT

A statistical and practical policy-oriented procedure for supporting or not supporting the use of student evaluations for faculty promotion decisions is described. The procedure is based on measures of teaching excellence. A comprehensive General Systems Theory type of flow diagram cross-classifies and causally relates all the relevant variables. Sources of invalidity for evaluation instruments are listed. Meaningful differences between ordinal scale measures are discussed, along with the need for statistical controls of intervening and extraneous variables. Thirteen questions that must be addressed to promote the fair use of student evaluation data for personnel purposes are raised. In addition, guidelines for college and university personnel policy are provided. It is recommended that validity studies be based on pre- and post-measures of amount learned by the student, independent of course examinations. The studies must control for all potentially significant intervening variables. Included are bibliographies on: the evaluation of teaching, alternate teaching approaches and problems in higher education, student evaluation of instruction and its validity, and legal aspects. (Author/SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED251053

LEARNING VERSUS RATING TEACHING

by Professor Carl Slawski

Sociology Department
California State University
Long Beach, California 90840

For presentation at The Conference on Postsecondary Education co-sponsored by Division J on Postsecondary Education of the American Educational Research Association (AERA) and The Association for the Study of Higher Education (ASHE).

Conference title: "Shaping the Future of American Higher Education: An Organizational Perspective".

Prepared for the section entitled:

"Faculty Futures: Systemic Factors Shaping the Academy"

Given in San Francisco,
Monday, October 29, 1984.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Carl
Slawski

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

HE017919

LEARNING VS. RATING TEACHING

by Carl Slawski

Sociology Department

California State University

Long Beach, CA 90840

Abstract

A statistical and practical policy-oriented procedure is presented and partly exemplified for supporting or not supporting faculty promotion decisions based upon measures of teaching excellence. A comprehensive General Systems Theory type of flow diagram cross-classifies and causally relates all the relevant variables. Sources of invalidity for evaluation instruments are listed. Meaningful differences between ordinal scale measures are discussed, along with the need for statistical controls of intervening and extraneous variables. Critical questions are raised about misuse of data. Positive guidelines are provided. Validity studies must be based on pre and post measures of amount learned by the student, independent of course examinations, and must control for all potentially significant intervening variables.

Faculty Perceptual Data

On Use of Student Evaluation of Instruction Data

For Personnel Purposes

In a survey reaching virtually all of about 800 full-time faculty members at a large public five-year college in the western United States in the Spring of 1982, a total of 276 responded to either a long or short-form (follow-up) questionnaire (a fairly good response of approximately 34.5%). Of the 179 responding to the question "How familiar are you with the research literature on validity of student evaluations for personnel purposes?" almost exactly the same number, 50% each, indicated that they rarely or seldom read anything on the subject 50 % indicated they at least occasionally read research articles addressing this issue. Only 7.3% indicated they were full conversant with the literature. This indicates a lack of knowledge of the subject on the part of most faculty members.

Of the 240 responding to the question "Are you satisfied with the official analyses and interpretations made of student evaluation data in the personnel process on this campus...by your department personnel committee?" 40% said "Yes", but 60% said "No" they were not satisfied. This would indicate a rather high degree of discontent with a mandated, campus-wide questionnaire to be used in at least two classes per semester by each full time faculty member.

Of 270 responding to the question "For you personally, do you believe that the use made of official student evaluation data

on this campus in the last three years or so has helped you to improve your teaching performance or skills?", (rounding to the nearest whole number) 29% indicated that it has helped them somewhat or a great deal, and the same number, 29% indicated that it has hindered the quality of their teaching either a great deal or somewhat. However, 42% said that it had little or no effect on their teaching. Thus, there was at least some constructive use made of these data by individual teachers.

Of 260 responding to the question "Overall do you believe that use made of student evaluation data on this campus in the last three years or so has aided the improvement of teaching performance?", only 23% indicated that it has helped either a great deal or somewhat. But 36% thought "It has had no visible or measurable effect" on teaching (in general). Adding the "no effect" to the "hindered" responses we see that more than three times as many respondents thought the use made of these data was either negative or neutral (74%) than thought the effect was positive (23%).

The judgment also appears negative when looking at the extreme categories on the four-point response scales. Namely, 10% thought the data helped a great deal, but 15% thought it hindered a great deal (overall). Similarly, when applied to their own specific case, 4% thought the data use helped a great deal, but more than two and a half times as many, 11% thought it hindered their teaching a great deal. This seems to indicate that more individuals are individually indifferent or at least

are able to cope with the data for their own analyses and interpretations than are willing to admit that its use might have had a positive effect on teaching overall on their campus. However, more seem to have been burned by it than have been strongly helped personally. While freely admitting the inability to calculate statistical significance of the data from this small survey for lack of a random sample, some strong indications can be stated.

The most startling statistic from this survey to the present writer is the fact that 74% thought the overall effect of the use made of these data was either negative or had no effect on teaching on this campus. Especially when no manual of interpretation is published by the administration for users of the data; when the limitations of the statistics are not published, such as confidence intervals (but only mean scores and percentiles are given); when the spurious effects of intervening or extraneous variables are not computed; when no warnings are given of the easy misuse of such data; and when no ongoing research on validity (as opposed to reliability) is being carried out, these facts bode ill for the fair and careful use of these data so highly prized by statistically unsophisticated administrators and personnel committee members. The conceptual model to be presented in this present paper then is a possible direction for some answers with powerful implications for a more even-handed evaluation of teaching, in a way that allows student input but maintains full faculty rights to due process.

Hypotheses to Accompany Diagram on

LEARNING VS. RATING TEACHING

The following hypotheses correspond to the numbers and letters on the diagram headed "Elements of LEARNING VS. RATING TEACHING". Notice that there are three columns in the diagram. The left column includes variables dealing with the "Raw Materials of Learning". The Second Column deals with the methods of learning and teaching, the processes involved, and facilities that ought to promote the learning process. The third column deals with both the immediate influences on the use of rating data as well as the outcomes or the ratings themselves. The diagram is also divided into two large rows of variables. The top row corresponds to elements or variables directly related to student background, attitude to learning and rating the teacher, as well as amount learned. The second row deals with teacher inputs in the form of methods employed in teaching, resources, as well as the ratings received by the teacher from students. This organization of the diagram is intended to help the researcher, counselor or data user to visualize the relationships between the variables in all their complexity.

<u>Key Measured or Possible Bases for Validity Inference</u>

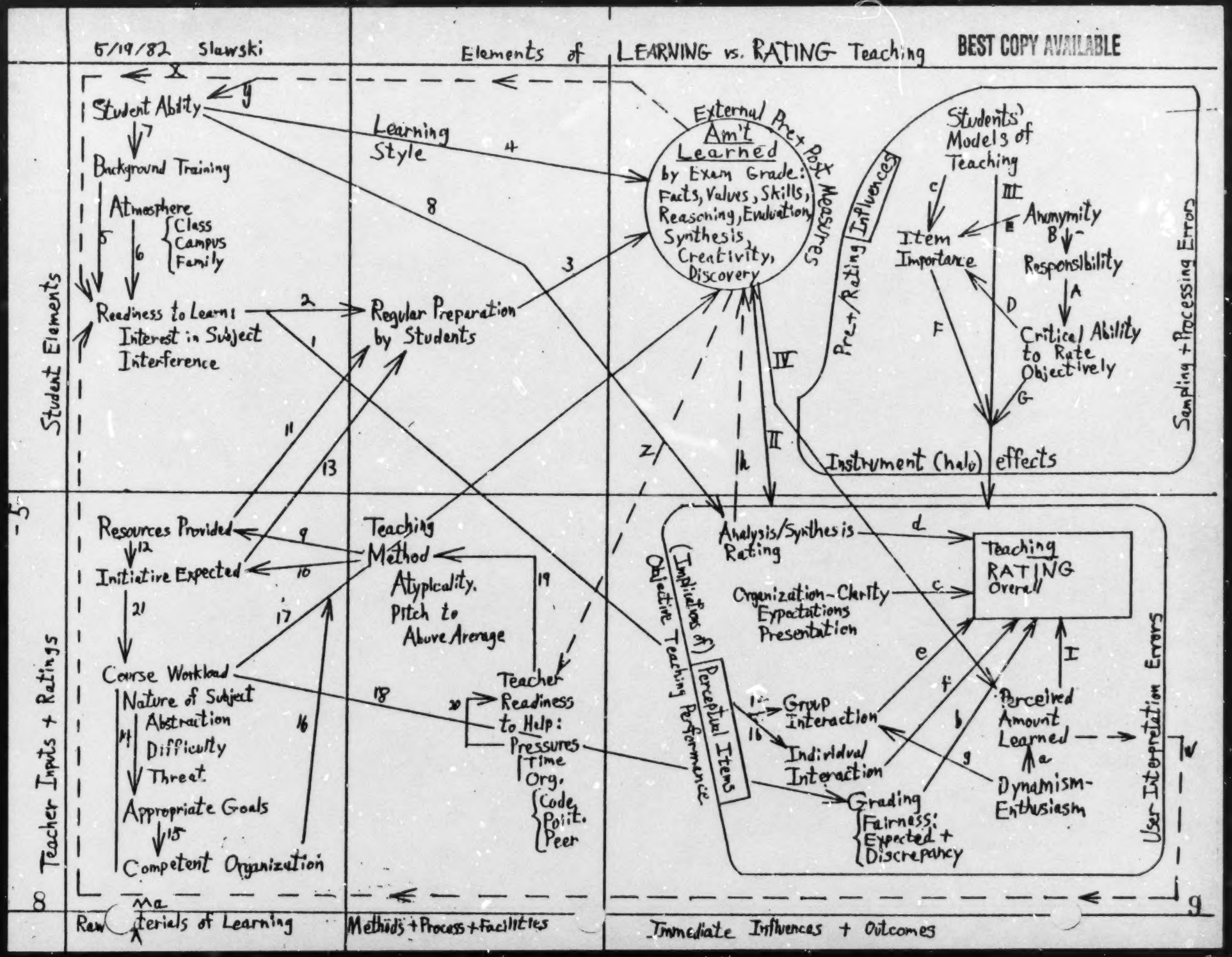
I. Perceived amount learned by students is directly related to overall rating given to teaching.

II. Amount learned by students, by objective pre and post measures, will affect future teaching performance.

BEST COPY AVAILABLE

5/19/82 Slawski

Elements of LEARNING vs. RATING Teaching



III. The students' models of teaching, as perceived by the teacher, will affect teaching performance.

IV. The amount actually learned (according to objective measures) directly affects perceived amount learned by students.

A. Multiple qualitative measures of deeper levels of learning are needed here.

Hypotheses Usually Tested in the Literature by Means of Perceptual Items and Correlations

a. Perceived dynamism or enthusiasm of the teacher is directly related to perceived amount learned by the student.

b. Perceived unfairness of grading practices (but especially the discrepancy between anticipated grade and that actually received) inversely affects the student's overall rating of the teacher.

c. Perceived organization and clarity of the teaching presentations (in relation to students' expectations) directly affects the overall rating of teaching received.

d. Students' perceptions of the quality of analysis and synthesis given to the subject matter directly affects their overall rating of teaching.

e. The quality and amount of group interaction perceived by students directly affects their overall rating of teaching.

f. The amount and quality of individual teacher-student interaction directly affects the overall rating of teaching.

g. The perceived dynamism and enthusiasm of the teacher

directly influences the quality of group interaction perceived by the students.

h. The perceived quality of analysis and synthesis presented by the teacher directly affects the amount learned by the students (by objective pre and post measures).

1) In order to decrease user interpretation errors, the above perceptual items need to be thoroughly supplemented if not supplanted by more objective observation by experts on teaching who are not part of the grading-rating game and the interaction process per se.

Instrument Effects and Other Methodological Biases

A. The responsibility of the student will influence his/her critical ability to rate a teacher objectively.

B. Anonymity of the rater (especially one who is a part of the teaching-learning process) will likely lead to less responsible ratings on the part of many (especially the disaffected) students.

C. A student's model of teaching will determine his/her judgment about the importance of a given rating item.

D. The student's critical ability to rate a teacher objectively will also help determine his/her judgment about the importance of a given rating item.

E. The fact of anonymity of a rater will likely skew the perceived relative importance of certain rating items.

F. The rater's judgment about the relative importance of an

item will influence his/her judgment about a particular rated teaching component.

G. The student's critical ability to rate a teacher objectively will affect his/her ratings on each content item.

Learning and Teaching Materials, Methods, Processes Plus Facilities

1. A student's readiness to learn is directly related to his/her perception of the quality of both a) group interaction and b) individual interaction with the teacher.
2. A student's readiness to learn will be directly related to regular preparation by the student.
3. Regular preparation will strongly influence the amount learned by the student.
4. A student's ability (and implicitly, his/her learning style) will directly affect the amount learned.
5. The student's background preparation or training will directly affect his/her readiness to learn.
6. The atmosphere of the classroom, campus, and the student's family situations will directly affect his/her readiness to learn.
7. A student's ability will correlate with his/her background preparation or training for the course.
8. A student's ability will correlate with his/her background preparation or training for the course.
9. The teaching method used will be directly related to the quality and amount of resources provided.

10. The teaching method used will directly affect the initiative expected of the students.
11. The amount of resources provided by the teacher promotes regular preparation by students.
12. The resources provided by the teacher directly affect the initiative expected by students.
13. The degree of initiative expected directly affects the regularity of preparation by students.
14. The nature of the subject (in terms of abstraction, difficulty or threat to the student) directly influences the appropriateness of course goals.
15. A statement of appropriate course goals will promote competent organization of the course content.
16. The level of competency of course organization, as mediated through the teaching method, affects the amount learned in direct proportion.
17. The course work load, as mediated through the teaching method, affects the amount learned in direct proportion.
18. The course work load directly affects the grading practices, perceived fairness and distribution.
19. The teacher's readiness to help determines or directly affects the method of instruction used.
20. Pressures on the teacher's time, as well as the formal and informal organization of the department and the school, limit the teacher's readiness to help the student. (Organization involves formal policy or the education,

political interaction, and peer interaction.)

21. The amount and kind of initiative expected of the student is the key factor in determining the student's work load for the course.

Feedback Processes

w. Ratings of teacher performance by students directly influence the student's subsequent readiness to learn.

x. The amount actually learned indirectly influences the student's readiness to learn.

y. The amount actually learned directly affects the student's ability to perform in the future.

z. The amount actually learned by students directly affects the teacher's readiness to help students.

The variables and hypotheses that ought to be measured with most diligence are hypotheses 2 and 15, plus pre and post measures of amount learned, quality of resources provided to students, and the teaching performance as measured by objective indicators and observers. The hypotheses that ought to be most central in research on the teaching-learning process include numbers 3, 11, 16 and 15. These judgments are based on the intuitive centrality of the just listed variables and hypotheses.

It is the administration's task to control or account for:

SOURCES OF INVALIDITY IN

STUDENT EVALUATION OF INSTRUCTION:

A COMPREHENSIVE CHECKLIST

RESEARCH DESIGN

Internal Validity = minimum control/treatment for interpretability

1. History -
2. Maturation -
3. Testing -
4. Instrumentation -
5. Statistical regression -
6. Selection -
7. Experimental mortality -
8. Selection-maturation interaction, etc. -

External Validity = representativeness or generalizability to other persons or groups.

9. Reactive or interaction effect of testing
10. Interaction effects of selection biases and the experimental variable -
11. Reactive effects of experimental arrangements
12. Multiple-treatment interference
    - A. Ecological fallacy
    - B. Negative impact studies to be done on:
        1) minorities, women

2) Alternate methods or styles of teaching

MEASUREMENT Validity

1. Face validity
2. Content validity -- to show job relatedness
   A. Meaning of rater vs. ambiguity of item or rater's understanding, imprecision, differences across the raters (such as readiness to learn, etc.).
      1) Linguistic and Material fallacies
3. Construct Validity
   A. Failure to triangulate
4. Criterion Validity
   A. Concurrent
   B. Predictive Validity
5. Meaningfulness of resulting data as presented:
   A. Ordinal level of measurement
   B. Statistical significance: Choice of technique
   C. Level of association
      1) Assumptions: specified distribution of sample, minimum frequency per cell, independent observations, known ratio of variances (or equality).

CONTROL FOR INTERVENING, Extraneous or Antecedent VARIABLES

1. 2 x 2 x 2 tables
   A. Significance between the difference between the difference between proportions.
2. List and justify intervening variables from the literature (with correlations and disagreements and alternate

interpretations).

A. Substantive Variables: e.g., Required course, work load, unusual methods, grades, transference relationship with instructor, T.A. games, subjective personality, etc.

B. Methodological Variables: e.g., response set, halo effect

1) Thirteen "Errors in Surveys"

C. Unusual substantive variables: Initiative expected, theoretical emphasis, innovative/unexpected teaching style or method.

On Errors in Surveys

The following lengthy quotation from W. Edwards Deming (American Sociological Review, August 1944) nicely summarizes the necessity of planning to control all extraneous or intervening variables in order to avoid "Errors in Surveys".

Reading 17 in Norman De (ed.), Sociological Methods, New York: McGraw-Hill, 1978, 239-247, quoting pp. 233-234.

"On Errors in Surveys"

by W. Edwards Deming

I. Purpose

There are thirteen different factors that affect the usefulness of surveys. The chief aim of this article is to point out the need for directing effort toward all of them when planning a survey, and the futility of concentrating on only one or two of them. Another aim is to point out the need for theories of bias and variability in response; such theories would correlate the vast amount of experience that is accumulating and would make comparability possible between different methods of collecting data. Biased methods could then be utilized in circumstances where they are more reliable and cheaper, just as biased samples are sometimes

used for gains in reliability.

In the planning of a survey, effort should be directed toward the reduction of all of the errors that it is possible to reduce, but the effort should be apportioned with a view to producing the greatest possible usefulness with the funds available. As a matter of fact, consideration of all the errors that will effect the ultimate usefulness of a survey largely determines whether it should be taken at all; and then, when the decision is made to go ahead, the same considerations should form the basis of planning. Burden of response, public relations, and availability of personnel and facilities are also governing factors.

There will also be an occasion in this article to point out

a) That the errors arising from some of the factors are larger than is commonly supposed;

b) The need for wider coverage and integrated research programs in the theoretical and empirical evaluations of some of the errors with the aim of (i) finding ways of reducing them; (ii) discovering the limitations that some of these errors place on the accuracy of a survey;

c) That precisions sometimes specified in the administrative requirements of a proposed survey are altogether unnecessary, and owing to noncompensating errors and biases have not in the past been realized;

d) That to eliminate sampling errors by taking a complete count will be a bad investment if other sources of error could be reduced and sufficient reliability attained at less cost;

e) That there is a special responsibility in presenting the data of a survey to give an account of all the difficulties of response and collection, in order that errors of interpretation and utilization may be reduced.

II. CLASSIFICATION OF FACTORS AFFECTING THE ULTIMATE USEFULNESS OF A SURVEY

The thirteen factors referred to are not always distinguishable and there are other ways of classifying them, but the following list has been found useful.

1. Variability in response;
2. Differences between different kinds and degrees of canvass;
   a. Mail, telephone, telegraph, direct interview;
   b. intensive vs. extensive interviews;

c. Long vs. short schedules;
d. Check block plan vs. response;
e. Correspondence panel and key reporters;
3. Bias and variation arising from the interviewer;
4. Bias of the auspices;
5. Imperfections in the design of the questionnaire and tabulation plans;
a. Lack of clarity in definitions; ambiguity; varying meanings of same word to different groups of people; eliciting an answer liable to misinterpretation;
b. Omitting questions that would be illuminating to the interpretation of other questions;
c. Emotionally toned words; leading questions; limiting response to a pattern;
d. Failing to perceive what tabulations would be most significant;
e. Encouraging nonresponse through formidable appearance;
6. Changes that take place in the universe before tabulations are available;
7. Bias arising from nonresponse (including omissions);
8. Bias arising from late reports;
9. Bias arising from an unrepresentative selection of date for the survey, or of the period covered;
10. Bias arising from an unrepresentative selection of respondents;
11. Sampling errors and biases;
12. Processing errors (coding, editing, calculating, tabulating, tallying, posting and consolidating);
13. Errors in interpretation;
a. Bias arising from bad curve fitting; wrong weighting; incorrect adjusting;
b. Misunderstanding the questionnaire; failure to take account of the respondents' difficulties (often through inadequate presentation of data); misunderstanding the method of collection and the nature of the data;
c. Personal bias in interpretation.

The foregoing factors will be ranked in importance one way for one type of survey and another way for another type. Thus, in a survey asking farmers their intentions to plant a certain crop next spring, the date of the survey would be especially important. For a survey of farm income the definition of farm income might easily be the most important factor.

Model Measuring Instrument

1. Behavior observable by untrained observer.

A. Focus on course experiences provided, not on instructor

personality/"performance".

2. Security of data collection procedures.
3. Admonitions of importance of whole and each item.
   A. Interpretation to raters of meaning of each item as intended by researcher.
4. Items focused on teachers' intended goals.
   A. Portion for personnel to measure course description and goals.
   B. Portion for student course choice to relate to grades, popularity, rapport, work load.
   C. Background items on students' perception of model teacher, student readiness to learn, interest in subject, reason for taking course, preparation, daily & prior, absenteeism.
5. Challengeable ratings:
   A. Peremptory: say 1 per each ten enrollees.
   B. Cause: reprisal, say for low expected grade, personality clash
      1) Due process here requires signed evaluations, especially for low ratings.
         a. Disclosure may wait until after grades are posted.
      2) Criteria for selection of "student jurors".
         a. Critical ability of observer/rater (including the interpreter).
      3) Dilemmas in evaluation of "facilitators".
   C. Disclosure of departmental (normative) raw data divided by appropriate subsample: lower division elective vs.

required; upper division, graduate seminar, even leaving out names and course numbers to maintain faculty confidentiality.

Manual of Interpretation and Justifying Decisions Based on These Data

1. Weighting of data to be limited to 10% to 25% of teaching component on promotion criteria or tenure, in relation to other evidence on teaching per se.
2. A full statement of possible sources of invalidity should be made to interpreters, promotion committee members each year, based on complete analysis.
   A. How complete is the sample? Is it "random" vs. "judgement/availability" type?
   B. Is the distribution normally distributed in relation to population? How known?
   C. What criteria remain unmet for statistical tests of significance or measures of association?
   D. What are the implicit theories or students for the teaching/learning experience, which may likely be unconscious?
      1) What are the implicit theories or teaching codes of interpreters/peers?
   E. What alternate explanations of the evaluation (say by the evaluatee him/herself) exist, and how are these accounted for, based on viable evidence?
   F. How are alternate explanations weighted in relation to a teaching code, if any?

G. How are data from an alternate instrument to be interpreted in relation to the more standard data collection instrument?

1) What administrative assistance is given, including funding to validate an alternative instrument? Or is there an administrative cop out, dereliction of duty to provide this aid to innovative, creative, or experimental teachers?

a. Shifting the burden of double evaluation and validation to the evaluatee who uses alternate methods or teaching styles is unfair.

H. Fallacies in inferences from data: Linguistic, Irrelevance, Material -- 28 types.

3. Peer consultation must be primary in interpreting data from students.

A. Ethics of authority, consultor's use of evidence (e.g., APA Ethical Standards).

4. Candidate must be told what specific aspects of teaching behavior are weak, promptly, in writing, in detail, with specific evidence, through discussion of alternate data interpretations, including the reasonability of demonstration of any administrator's judgment which is contrary to that of peers, if any. Otherwise there is no fair criterion for improvement, but space for arbitrariness.

Meaningful Differences Between Ordinal Scale Measures

The following simple and largely intuitive statement

describes the underlying reasons why ordinal scales and median scores are appropriate when deciding differences between two or more instructors' ratings. Perceptual questionnaire data can never reach the interval level of measurement where mean scores would be appropriate according to the traditional mathematical assumptions so easily violated by naive social scientists and educators.

The basic data from student evaluation forms has always been on a scale of whole numbers, e.g., 1, 2, 3, 4, 5, etc. Thus there exists an ordinal level of measurement at best. For illustrative purposes let us assume a five-point scale, such as the following.

```
I..'..'..'..'..I..'..'..'..'..I..'..'..'..'..I..'..'..'..'..I
5               4               3               2               1
Excellent    Above          Average         Below          Poor
             Average                        Average
```

Assume an odd number of students in each class for the individual instructor as well as an odd number in all classes in the department for a given semester.

Hypothetical example: For a given instructor on a given questionnaire item, for the sake of simplicity, assume a sample of seven students rated his/her performance, falling in an array such as this: 5 4 4 3 3 3 1

The median is the middle number, a "3".

The mean would be 23/7 or 3.3 in round numbers.

This measure using the mean is mathematically illegitimate, since the mean assumes at least an interval scale or level of

measurement, and the basic data is only ordinal. To reach an interval level, the basic data would be such that the difference between 1 (= Poor) and 2 (= Below average) is exactly equal to the difference between 4 (= Above Average) and 5 (= Excellent). And of course there is no way to measure these differences.

The department mean, if complete raw data were available to compute it (which it is not), would likely be between 3 and 4, perhaps at 3.5. If so, the candidate/instructor would be 0.2 below this unvalidated "norm".

However, if one student of that instructor changed his (the student's) rating from a 3 to a 4, that would give the array: 5 4 4 4 3 3 1, and the median for that instructor's class would now be a "4".

On a purely intuitive basis, one student or one rater changing a rating of only one point on this item could not constitute a meaningful or significant difference. A meaningful mathematical or statistical difference between one median and another would then have to be something greater than "1", namely a difference of at least "2".

Because in practice department medians will nearly always be either a 3 or 4, the candidate's median, to be considered meaningfully below the department median for a given course would have to be either a 2 or a "1" on the five-point scale. This rarely happens. It certainly has never happened for any such data I have seen.

The overall conclusion would seem to be that for personnel purposes, differences in mean scores, or even median scores ought

to be ignored, except in extreme cases, or for very wide differences. Measures like the standard error of measurement ought to be calculated routinely, and preferably a weighted confidence interval (weighted for validity and intervening substantive variables, such as work load, grade expected, orthodoxy of teaching style, etc.). See separate explanation of confidence interval.

The Need for Confidence Intervals

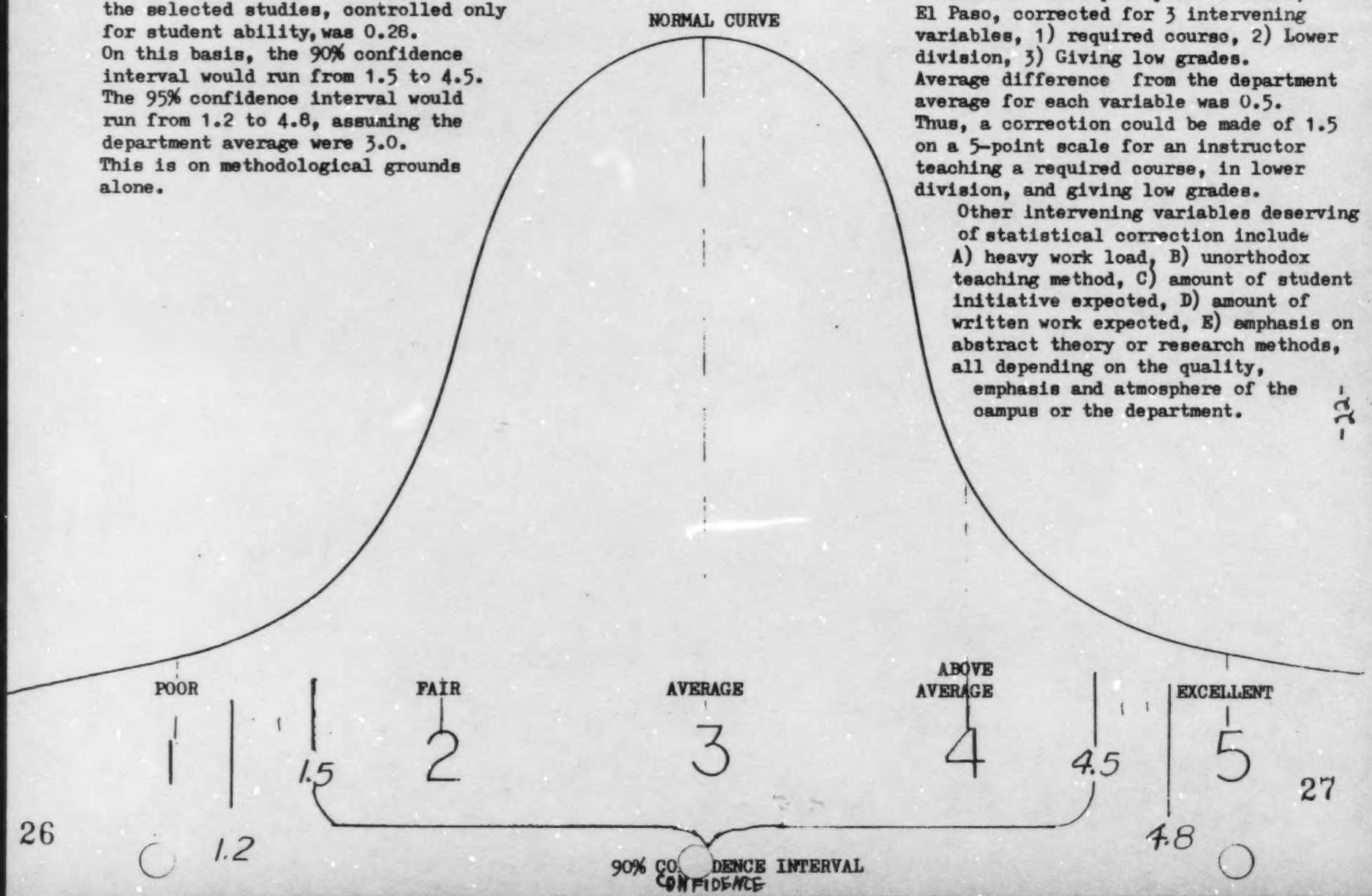
The "Normal Curve" diagram illustrates the use of the confidence interval and the consequent extreme differences required between ratings of instructors in order to show a mathematically or methodologically significant difference between them. These large differences hold even before considering substantive differences between instructors, teaching methods and styles, student interest or readiness to learn, etc. The baseline correlations come from research analyses reported by Dowell and Neal (1982).

William G. Cochran's posthumous methods book, entitled Planning and Analysis of Observational Studies (Wiley, 1983), notes the unfortunate pattern common to social scientists looking for significant differences rather than predicting the specific value a variable should take on theoretical or law-based grounds. This would seem to support the present author's preference for confidence intervals and measures of predictive validity rather than the common naive use of percentiles based on differences between mean scores between a candidate/instructor and his

Slawski, C., 11/18/83

Dowell, D., & J. Neal, "A Selective Review of the Validity of Student Ratings of Teaching," J. Higher Education 1982, 53, 1, 51-62. See especially pg. 58 on confidence intervals.

The median validity correlation for the selected studies, controlled only for student ability, was 0.28. On this basis, the 90% confidence interval would run from 1.5 to 4.5. The 95% confidence interval would run from 1.2 to 4.8, assuming the department average were 3.0. This is on methodological grounds alone.



A research based policy at U. Texas, El Paso, corrected for 3 intervening variables, 1) required course, 2) Lower division, 3) Giving low grades. Average difference from the department average for each variable was 0.5. Thus, a correction could be made of 1.5 on a 5-point scale for an instructor teaching a required course, in lower division, and giving low grades.

Other intervening variables deserving of statistical correction include A) heavy work load, B) unorthodox teaching method, C) amount of student initiative expected, D) amount of written work expected, E) emphasis on abstract theory or research methods, all depending on the quality, emphasis and atmosphere of the campus or the department.

department average. Such a more rational process may well foretell the need to devalue the emphasis given to such quantitative student evaluation data in the personnel process, at least until much more highly sophisticated theories and methods for their use and especially for their interpretation are developed and refined.

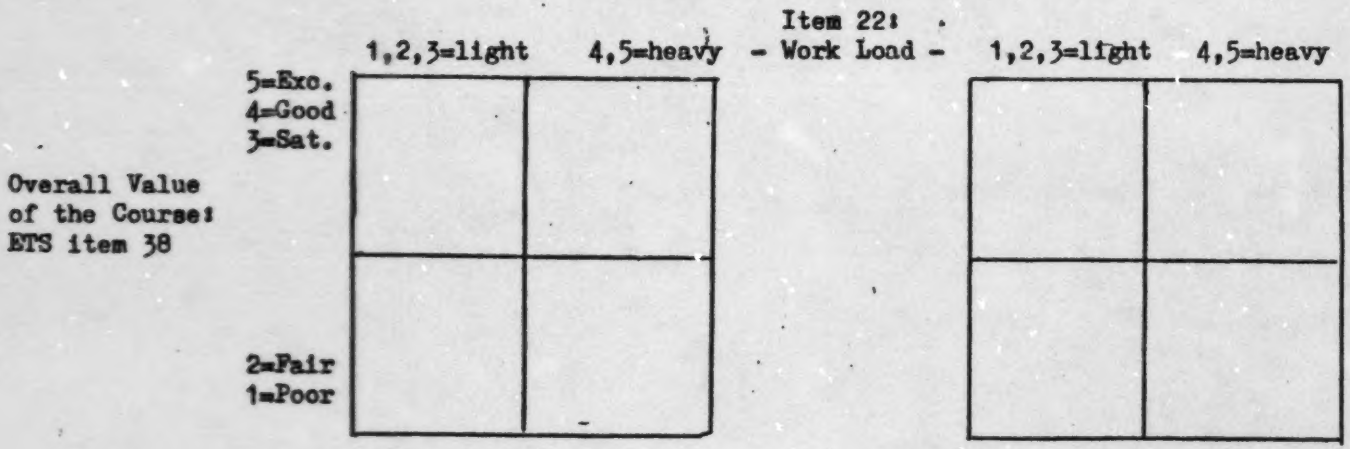
Employing Statistical Controls

The following 2 x 2 x 2 table and commentary illustrates the necessity of procuring the raw data from each student's questionnaire. It is necessary in order to control for each possible intervening or extraneous variable while correlating the desired main independent and dependent variables as is shown to be possible by filling in these dummy 2 x 2 x 2 tables. The description assumes for illustrative purposes that the ETS-SIR forms were used to rate the instructor.

2X2X2 CONTINGENCY TABLES

Instructor A

Whole Department



Item 38: I would rate the overall value of this course to me as: 5=Excellent, to 1=Poor
Item 22: The work load for this course in relation to other courses of equal credit was:
For item 22, responses on ETS form are: 1=much lighter; 2=lighter; 3=About the same; 4=heavier; 5=Much Heavier

This table cannot be completed unless the raw data were available from ETS for instructor A as well as for every instructor rated for the comparison group, namely the whole department in a given semester. These data have never been available from ETS for this campus. So any inferences made about the relationship between these three variables, 1) work load, 2) rating of overall value of the course, and 3) instructor in relation to the whole department average, could not be tested. It would have to be an educated guess, based perhaps on discussions with students in instructor A's classes or student letters of recommendation. In other words, the data provided to the department and to the individual instructor from ETS does not allow statistical control of the variables operating. No research can be done from these data alone to judge the effect of work load on the students' overall rating of the course either for the individual instructor or for the whole department. The same is true of any of the other among the 39 items of the ETS rating form.

Statistical Tests on Actual Ratings

The following analysis is based on data gathered through the use of the Educational Testing Service SIR (Student Instructional Report) form for the author's mean scores for four semesters, from Fall 1978 through Spring 1980. (Because the data is from the author's own classes, it seems more appropriate to revert in this section to the first person singular.) Focus here will be on the two summary items from the ten classes evaluated, namely, items 38 and 39. Item 38 rates the "overall value of the course", and is the item I believe was most appropriate because of the student-centered teaching method as an overall measure for student ratings of the course. Item 39 rates "teaching effectiveness". I realize that the mean is not appropriate for ordinal data, but since the campus administration insisted that mean scores and percentiles be used, I decided to go along with the assumption that these measures were valid not to mention adequate for making personnel decisions. Carrying the same administrative logic to its ultimate conclusion on mathematical grounds, I performed the following simple statistical tests on the data for my classes in comparison to the mean scores for my department as a whole for each of the corresponding semesters' data.

In order to find out whether my ETS mean scores were different from or above or below the departmental average by a statistically significant amount, I performed t-tests, using the standard formula, as follows:

$$t = \frac{X_{me} - X_{dept.}}{S_{dept.}/\sqrt{n}}, \text{ where } S_{dept.} = \frac{\sum_{1=1}^{N} (Xi - X)^2}{N}$$

1) Item 39, My Just Below the Median of Mean Scores

For the course in which my mean score on item 39 was just-below-the median of my ten mean scores for that item, my mean was 2.7. The department mean was 3.5. There were eleven students who rated that course (Sociology 456, in Fall 1978). The Standard Deviation for the department was S = 1.1.

Given the assumptions that are violated by using the t-test here, such as sampling bias, lack of normal distribution, lack of interval level of measurement, etc., it is reasonable to make the most conservative use of this t-test, namely by using the level of significance of = 0.01, and a two-tail test. Due to small sample size, the t distribution is only approximate.

The actual t-test produced a value of 2.41. For a two-tail test, at the 0.01 level, with d.f. (degrees of freedom) = 11-1 = 10, the table value = 3.169. This is larger than the actual value, so we cannot reject the null hypothesis. Thus, there is no statistically significant evidence of a difference between my scores on this item 39 for Sociology 336 for Fall 1978, and the department average. Furthermore, since this course is the one in which my mean score is just below the median of all ten courses rated, I also might conclude that there is likely no difference overall between my averages for this item (#39) in the ten rated classes and the averages for my department for the four semesters

at issue.

2) Item 38, Just Below the Median

I then performed the same test for my just-below-the median of mean scores for the other global item, item 38 (overall value of the course). That occurred in Fall of 1978, for Sociology 336 (Small Groups, the first time I taught the course). My mean was 3.18. The department mean was 3.66. The Standard Deviation for the Department was = S = 1.18. There were seventeen rating students in the class. The t-test produced a value of 1.68. For two-tail test, with d.f. = 16, at the 0.01 level of significance, there was again no significant difference between my score and the department for that course. Thus, we cannot reject the null hypothesis, and there is no likely difference between my averages for item 38 and for the department averages for that item for that course (Sociology 456) as well as for the ten courses in the four semesters at issue. I have always contended that 38 is more appropriate for my teaching method than is item 39.

3) Low Rating on Item 38

I proceeded to compute the same t-test on item 38 in the semester in which I received the lowest rating on that item (excepting only the two relatively new courses taught in the term in which I lost my voice for an entire month). This next to the lowest rating occurred in Sociology 456, Fall 1978. There were 11 rating students, who gave me a mean rating of 2.7. The department mean was 3.66. Standard Deviation for the department was = S = 1.38. The t-test produced a value of 2.31. For a two-

tail test at the 0.01 level of significance, with d.f. = 10, the null hypothesis of no difference again cannot be rejected. Even on this next to the lowest rating (among my ten courses rated), there is no significant difference between my scores and the department average.

4) Item 38, Missing High Rating

I tried the same t-test on the data not included in my file, that for Sociology 335, Social Psychology, Fall 1979, item 38. I discovered first of all that my mean score for that course was higher than for any other course rated by students in the four semesters and ten courses at issue. My mean was 4.20 (on a scale of 5 = high). There were 15 rating students. The department mean was 3.7. Standard Deviation for the department was = S = 1.05. The t-test produced a value of 1.84. For the ordinarly two-tail test at the 0.01 level of significance, this value would not allow us to reject the null hypothesis.

Chi Square

Next I arranged the data for the same items and the same four courses as above, for items 38 and 39, in order to perform a chi square test using a 2x2 contingency table comparing my scores to those of the department. I dichotomized the data into high scores (3, 4, or 5) and low scores (1 or 2) for each item, and employed Siegel's (Nonparametric Statistics, 1956, pg. 109) formula 6.4, corrected for continuity (since the sample is greater than 40). Since this test does not require interval data, chi square is more appropriate than the t-test. However,

almost the same basic results were found. Note that one of the four cells in each test had an expected frequency of less than 5, so technically this test is not appropriate.

1) For my just-below-the-median of mean scores on item 39, the chi square value was 2.265. From this result we cannot reject the null hypothesis. For this more appropriate test (which is thus more sensitive to extraneous variables or false positive correlations) it is reasonable to use the ordinarily accepted 0.05 level of significance instead of the 0.01 (as used above for the t-test). The two-tail table value for d.f. = 1, at the 0.05 level, is 3.84. Thus there is again no statistical evidence that my scores for item 39 are significantly different than those of the department (on "teacher effectiveness"). This confirms the results of the t-test above).

2) For my just-below-the-median of mean scores for item 38, the chi square value obtained was 2.156. From this we cannot reject the null hypothesis of no difference at the 0.05 level for a two-tail test. Thus there is no statistical evidence that my scores are different from the department, for this course (Soc. 336, Fall 1978) or overall, among the ten courses rated in the four semesters at issue. This once again confirms the results of the t-test above.

3) For my next to the lowest score on item 38 (as above), in Sociology 456, Fall 1978, the chi square value obtained was 1.287. For a two-tail test, even at the more restrictive 0.01 level, again we cannot reject the null hypothesis. There is no

evidence here that my scores are statistically different from the department average. The t-test results are again confirmed.

4) My highest rating on item 38 (the one omitted from Table II) was for Sociology 335, Fall 1979. The chi square value obtained was 0.982. We cannot reject the null hypothesis. There is no evidence that my scores are different from the department scores. This again confirms the findings of the t-test performed above.

Finally, I performed a Pearson's r, correlation for item 38 between my mean scores and department means across the ten rated courses at issue, and found a correlation of 0.925. This is a very high correlation, and good evidence that on this global item ("overall value of the course") my scores are predictable from those of the department. Or in other words, my scores are given virtually parallel value with those of the department.

Notice that this entire analysis does not take account of, nor control for the effects of any intervening or extraneous variables, the serious, ever-present sampling errors, the assumptions for statistical tests (such as equal variance of the populations), let alone the validity of measurement, or the numerous possible errors in surveys (especially the thirteen types of error discussed elsewhere in my file). Nor does it even take account of the validity studies and confidence interval calculations of Dowell and Neal's article (Journal of Higher Education, 1982). If these many additional variables were to be taken into account, there would be even less possibility of any

statistically significant differences between my scores and those of the department.

The next section raises many of the questions that must be investigated and/or answered in order to be able to forward the fair use of student evaluation data for personnel purposes. It is the duty of administrators who use such data to sponsor the necessary research as a prerequisite to using it for such important decisions as tenure or promotion.

LEARNING VERSUS RATING TEACHING:

Outline of Questions and Some Answers for a Technical Evalation Manifesto

In general, these recommendations follow the principles expounded in the American Psychological Association joint Standards for Educational and Psychological Tests as well as a thorough review of the research literature, in addition to an understanding of administrative law related to use of test material for personnel purposes.

1. What specific alternate evaluation instruments (from published research studies) have given the same or equivalent ratings to the items dealing with satisfaction with teaching or with the course? Such equivalent instruments would be necessary to demonstrate validity according to the highest, written professional standards (as embodied in the APA STANDARDS).

2. How do you know (or what research studies have shown) that students nearly always know when they are learning a subject? On the contrary, if they want to learn immediately

useful career skills (a dominant current malady), but are required to learn humanistic or appreciative values, or the meaning of say chi-square or confidence intervals, they are likely to be dissatisfied with the course.

3. The use of items not directed carefully at the stated teaching objectives for each specific course gives only suggestive results at best and is minimally meaningful, let alone useful or valid for promotion purposes.

4. Is there any campus-wide evaluation system that has been structured to include measures of deeper levels of learning? This would include measures of such concepts as increases in empathic ability, discussion skills, insight facilitation, creative generation or use of novel ideas or procedures in practice, small group leadership, logical reasoning, aesthetic appreciation, making ethical judgments, etc.

5. On any single item, such as the "overall effectiveness" item, distinctions between instructors in a given department which are finer than classifying them into the upper, middle, or lower thirds, are probably not tenable or meaningful given the present state of the art of measurement in survey research on student evaluation of instruction.

6. What kinds of teachers are likely to be underrated by students, or to receive bimodal ratings? In a traditional or typical larger public college or department (oriented primarily toward teaching rather than toward research or doctoral education), underrated teachers are more likely to be those who

employ novel teaching methods, the more demanding in terms of student work load, the soft-spoken, non-traditional, visionary, politically deviant or ideologically eccentric, or more theoretical in orientation.

7. Assuming equivalent and valid sampling of classes (rare or non-existent on a large scale, e.g., campus-wide), results in terms of a mean or median that do not show a statistically significant difference from a departmental average cannot validity or ethically be used as the primary measure to demonstrate that an instructor is less or more effective than the average instructor in that department. Requirements such as random sampling, proper level of measurement, attainment of the normal curve, etc., are rarely met.

8. If such data are to be used for (any critical decision, such as for) promotion purposes, failure to use "confidence intervals" when comparing instructors' scores across courses is a grave deficiency, and may very well render such comparisons invalid.

9. Is this not a fair statement? A validity correlation of say 0.28 or 0.5 for a given item on a student evaluation questionnaire means that the item in question may be a valid predictor of the amount students learn only 28% or 50% of the time? If so, who would dare to make such data the overriding factor in any single personnel decision?

10. Would not true validity require measures of actual amount learned, as opposed to mere "perceived" learning, let alone

"satisfaction" with the course? In comparing 1) "satisfaction with course" and 2) pre and post measures of amount learned independent of course examinations, is a correlation of 0.5 enough?

11. If the connection is wrong half of the time (between course satisfaction and amount learned), can these kinds of perceptual data be used (say without confidence intervals to take account of "margins of error") in order to judge small differences between teaching effectiveness when comparing two teachers? 'YES' for guidance (or formative) purposes; 'NO' for promotion (or summative) purposes.

12. Has this kind of care (namely, 1) use of alternate instruments; 2) careful control of intervening variables; 3) measures of amount learned, and 4) use of confidence intervals) been accomplished as a matter of course in any campus-wide, on-going evaluation system? If so, where? And where are the results published?

13. What are the weaknesses of meta-analyses? Each study controls different variables. Rarely do the specific studies in the selected list measure amount learned, use alternate instruments or confidence intervals, or control across classes or individual learners for other important intervening or extraneous variables. Meta-analyses depend on published studies which are biased by the journal review process toward studies with positive findings. Thus meta-analysis produces and even compounds spurious correlations or inaccurate conclusions.

SUMMARY: Policy and Administrative Considerations
For a Better Future in Higher Education

For the following reasons it is essential to apply the highest possible professional standards of research to the use of quantitative student evaluation data for faculty personnel purpose.

1. The use of anonymous, quantitative is not an assurance nor a way of promoting the precedence of teaching over research activities of faculty (as is sometimes promoted).

2. Academic freedom may be endangered when A) students rate a teacher low because s/he teaches unpopular political views, or B) students rate the teacher low because s/he uses teaching methods that are unorthodox (though more effective) simply because the methods are inconvenient or students are unfamiliar with the teacher's methods.

3. The Code of Federal Regulations (29, Labor, parts 900 to 1899, revised as of July 1, 1981, part 1607) summarizing the "Uniform Guidelines on Employee Selection Procedures" (1978) calls for application of the guidelines by the Equal Employment Opportunity Commission in the enforcement of title VII of the Civil Rights Act of 1964, dealing with discrimination in employment. The Code refers to standards for validity studies as embodied in the Standards for Educational and Psychological Tests (prepared by a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education). In other

words, these APA standards should be applied to the preparation and use of quantitative student evaluation data in stringent fashion. Most importantly, the Standards call for ongoing "predictive validity" studies which would require correlations of the ratings of faculty with amount learned by the students. It would also mandate alternate evaluation forms, computation of confidence intervals, and statistical control for possible intervening or extraneous variables for each candidate for tenure or promotion. It also mandates many other carefully applied procedures of data collection and above all of their interpretation by the users at all levels of use (i.e., in the process of recommendation and decision).

4. These data (especially the quantitative data) are particularly subject to "arbitrary and capricious" use by recommending committees and administrators, in part because the numbers are so seductive and easy to twist to a premade judgment about the merits of the candidate. For example, it would be easy in practice, in the process of comparing data from fifteen or more candidates, to decide to promote those whose scores are above the mean of their department, and to decide not to promote those whose scores are below the mean. The mean of course is invalid for such ordinal level data. Such use is clearly "arbitrary and capricious" and against all the canons of objective, social scientific research. It is common, and an easy error consistent with the need for more universalistic standards. However, it is an especially grievous error when the quantitative

data are in practice, or de facto, the primary or exclusive basis for the decision of whether or not to promote or grant tenure.

It has been observed that faculty committee members and administrators often decide on the promotability of a candidate on gut feeling (or even on political or other bases irrelevant to competence), then justify the decision on the basis of whatever data is available in the file, such as the easily manipulable student evaluation data as the primary or dominant type of information.

5. Primary responsibility for the <u>interpretation</u> of information on teaching effectiveness ought to be in the hands of departmental peers. Discrepancies in interpretation ought to be resolved by the final administrator in writing, in keeping with precedents set by the U.S. Office of Personnel Management, the Administrative Procedures Act, as well as many college policies and guidelines and resolutions of faculty unions.

Legally, the president of each campus (in most instances) is responsible for the final decision, and thus also is directly responsible for explaining a decision not to promote or grant tenure, and to do it in writing.

6. It is the faculty member's legal, even constitutional right to have any negative decision explained in writing in signed documents. The candidate has a right by the fifth amendment not to be "deprived of...property, without due process of law". Promotion is a question of property rights, though tenure decisions are more critical in the eyes of the courts.

A. The sixth amendment, in the clause reading "the right... to be confronted with the witnesses against him" can easily be extended to mean that basing a negative personnel decision on the basis of anonymous student evaluation questionnaires by say three to five students (which is all it takes to get lower than department average ratings), is unconstitutional. It denies one the right to confront his/her accuser. Protection of students might be achieved by having them sign their evaluation, but assure that the results will not be seen by the instructor until after grades have been recorded by the registrar, or possibly in the case of a grievance initiative.

B. The legal notion of "substantive due process" means that faculty rights should be protected by written policies and written explanations for negative decisions, and that the data used as the basis for the decision should be contained in a file available to the candidate, and be the sole basis, along with public oral testimony, for making promotion or tenure decisions in the matter of teaching effectiveness. This means that the final administrator making a negative decision must not merely cite the policy on his legal obligation and state that his negative decision was based upon information contained in "the file", but rather on explicit statements about his interpretation of the meaning of the data and other evidence in the file. This means following the spirit as well as the letter of the law and policy.

7. Applied social science faculty should set a model and be

on the cutting edge of excellent evaluation research for such a key issue affecting their careers, their very demonstration of competence as social scientists or educators, their institutional well being, and their profession, especially as these decisions affect the newer, tenure-track faculty in their field. They should work to supersede the present de facto models of deplorable superficiality of data collection, above all of superficiality regarding analysis and interpretation of data in decision-influencing committees.

Positive precedents or models of social science evaluation research must ultimately show us how to synthesize qualitative with quantitative data. Until such models of research and procedures for data use are developed and widely implemented in colleges, the best prior examples of evaluation research for educational testing and public policy should in turn be the prototypes for collection and interpretation of these data. For the use of these data are absolutely critical to the future of individual careers and to the future of our professions as educators and social scientists.

FORMATIVE QUESTIONS

The following questions, or ones like them, ought to be asked at some point in the process for formative purposes, i.e., to help improve teaching. But because of the likely subjectivity of possible student responses it would not be wise to use such qualitative data for personnel decisions except at the discretion of the candidate/instructor.

To be Asked at Mid-Term

In order to help improve a course, questions like the following might be asked of students about mid-term, followed by small group discussions among students leading to practical suggestions to the instructor. Take about 15 minutes to answer the following five questions.

I. What do you like most about this course?

II. What do you like least about this course?

III. What would you like most to change about this course?

IV. Considering the course description and stated objectives of the course, what practical suggestions do you have to improve the course for the remainder of this term?

V. How, practically speaking, step-by-step, and in as much detail as possible, would you suggest implementing your suggested improvement(s)?

A.

B.

C.

Form groups of six to eight students and discuss your suggestions for improving the course. Then offer your suggestions to the discussion leader or coordinator (such as the director of the faculty development center or some respected facilitator).

Then the coordinator will present the class suggestions and options to the instructor, with an opportunity for the instructor to respond and discuss the suggestions with the class in the presence of the coordinator.

Use or implementation of the suggestions will remain up to the instructor. This information will remain confidential unless the instructor chooses to reveal it or even use it in personnel deliberations.

Use a five-point scale with space for comments.

Formative, supplementary questions available for a file open to students or the personnel committees at the option of the instructor:

1. Reason for taking this course: major requirement, elective outside major, etc.
2. How difficult were the assignments?
3. What grade do you expect to receive?
4. How well did you like the instructor?
5. On the average, how many hours per week did you study or work in preparation for the class meetings?
6. How much do you feel you learned in comparison to other courses in this general subject matter area?
7. Did the instructor hold to announced office hours?
8. How many times did you visit the instructor in office hours for five minutes or more during the term?
9. Did the instructor allow student questions during class time?
10. Were alternate tasks or options available for completing some of the course requirements?
11. How much student initiative was required or encouraged by the course objectives?
    A. How interested are you in becoming an independent, active learner?

12. The overall value of the course to me was:
13. How interested are you now in the subject matter of this course?
14. Have the course experiences provided you with the opportunity to exercise and develop your higher faculties, such as analysis, synthesis, evaluation, empathy, application to practical problems, or the generation of practical solutions?

Optional items (to be answered on a third punch card):

1. Did the format of the class allow or encourage you to learn from your fellow students, either in small groups or in other ways?
2. Did course experiences give you the opportunity to better communicate your values and feelings to others.
3. Has the course provided me the opportunity or tools to allow me to explore or enhance my sense of self identity?
4. Has the course allowed me the opportunity to facilitate another person's growth?

Any publicly available summary of student responses, to such questions should be preceded by a catalog description of the course, and a statement of the teacher's goals and responses, if any, to the evaluations, at the instructor's option.

Demographic items should also be included, such as student's major, minor, age, sex, etc., year in college.

Humanistic/Qualitative Depth Questions

The following quotations from John P. Miller's Humanizing

the Classroom (1976, pp. 172-173) suggest some very thoughtful additional questions that might be used for formative purposes. They are subject to strong elements of subjective interpretation and should not be required of all faculty or used for personnel purposes without explicit permission of the instructor being rated.

It is also possible to involve students in the evaluation process. This can lead to shared expectations between the students and teacher concerning curriculum outcomes. For example, the teacher and students could develop a set of questions that could be used to evaluate the students' development. Questions that students and teacher could examine together might include:

Can I function effectively in a group?

Do I hinder or facilitate the group process?

Do I give and openly receive feedback?

Do I make personal value decisions and do I commit myself to those decisions?

Do I support other students' efforts to make responsible decisions?

Am I able to communicate empathy, regard, and genuineness?

In what ways am I moving toward becoming a fully functioning person? In what ways am I not?

Am I becoming more centered and calm?

Am I developing my Higher Self? How?

Am I dis-identifying from personality models that are

restrictive to my development?

In what ways am I more creative and imaginative? In what areas can I improve my imagination and creativity?

What values are central to my sense of self? Do I act consistently on those values? What values would I like to develop and realize?

What are my principal concerns? How do these concerns relate to my sense of self identity? In what ways can I deal with these concerns in order to realize a positive identity?

Am I living in the here and now? Am I able to relate my thoughts to my feelings? Do I see things holistically or in segments?

Am I committed to growth and development? Am I willing to take risks or do I avoid opportunities for growth? Do I facilitate another person's growth and development? If so, how? If not, why not?

The final list of questions would depend on models used in the classroom.

SUGGESTED POLICY:

Qualifying Quantitative Student Evaluation of Instruction

I believe the following guidelines should be implemented in the personnel policies of colleges and universities. They are stated in approximate chronological order of importance, starting with those most easily able to be introduced or implemented.

1. If policy requires employment of a mandatory student evaluation questionnaire, limit the required number of courses

per instructor to be evaluated to two per year. The choice of courses should be made by the instructor by the fourth week of classes.

2. Limit the number of items required to be used for personnel purposes from two to a maximum of six global and behaviorally observable items clearly within the scope of students' abilities to make valid observations. These types of items should supersede the more purely perceptual and items of subjective opinion.

A. Each instructor should be allowed to select other items or qualitatively different forms of evaluation to be used at his/her discretion for personnel purposes.

B. Whole departments should be allowed to add additional mandatory items only by unanimous consent of all full time, tenure-track faculty members.

C. Items may be open ended, focused or fixed-choice, as seems most appropriate.

3. Insist that every item employed be subjected to thorough studies of their predictive validity in relation to the ultimate criterion of objective pre and post measures of amount learned in a course. These measures must be independent of course examinations unless such examinations are mandatory for all those teaching the same course in the department.

A. Unless impossible, measures of amount learned should include learning in terms of facts, skills, cognitive reasoning and judgment, practical application of knowledge

gained to settings beyond the classroom, or in other words, of higher levels of learning, of potentially lasting benefit.

B. Measures of amount learned should as far as possible be made in comparison to the same course taught in different semesters and by other instructors.

C. Validity studies, including content, construct, and predictive (or criterion) validity must be accomplished before such items may be used for personnel purposes. Studies on other campuses will not be considered sufficient to justify validity on a given campus.

4. The maximum weighting of student evaluation questionnaire data in relation to other evidence and testimony on teaching effectiveness shall be limited to between ten percent and twenty five percent, so far as the promotion or tenure process is concerned.

A. Continuing studies, financed by the administration, and with participation of interested research oriented faculty member, shall be done of various methods of evaluating teaching effectiveness and amount learned, together with released time to average one quarter time position for every twenty faculty members. Each interested faculty member will participate in rotation.

B. Each School within a university campus shall fund a half-time, faculty-elected teaching mentor, outside the personnel process.

5. Interpretation of student evaluation data shall require

controlled statistical comparisons of each candidate's scores against data for the whole department, controlling for extraneous or intervening variables, as far as they have been measured.

A. This can be accomplished without revealing the names of any instructor or specific course number, through computerized manipulation.

B. A wide range of questions should be addressed to every student on such instruments to inquire whether s/he has kept up with the assigned readings, spent how many hours studying each week, whether s/he is taking the course because it is required, his/her attitude toward the importance of the subject matter, the theoretical level, attendance, and background preparation, work load as s/he sees it, etc. This guidelines recognizes that the main burden of learning is on the student.

C. Forms should be signed or identifiable (after grades are submitted) to allow effective constitutional challenge of possibly unfair low ratings.

BIBLIOGRAPHY ON EVALUATION OF TEACHING

I. Statistics, Methods of Research, Quantitative and Qualitative Data Analysis and Interpretation

American Psychological Association and the American Educational Research Association and the National Council on Measurement in Education
1974 Standards for Educational and Psychological Tests. Washington, D.C.: American Psychological Association.

Andrews, Frank M., et al
1974 "A Guide for Selecting Statistical Techniques for Analyzing Social Science Data." Ann Arbor: The University of Michigan Survey Research Center.

Argyris, Chris
1980 Inner Contradictions of Rigorous Research. New York: Academic Press.

Blalock, Hubert M., Jr.
1960 Social Statistics. New York: McGraw-Hill.

Campbell, Donald T., and Donald W. Fiske
1959 "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." Psychological Bulletin 56: 81-105.

Campbell, Donald T., and Julian C. Stanley
1963 Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally. Reprinted from Handbook of Research on Teaching.

Chaterjee, Samprit, and Betram Price
1977 Regression Analysis by Example. New York: Wiley.

Coleman, James S.
1972 "Policy Research in the Social Sciences." Morristown, N.J.: General Learning Press. 23 p.

Cook, Thomas D., and Donald T. Campbell
1979 Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally. Ch. 2: Validity, 37-146.

Costner, Herbert L., and L. Wesley Wagner
1965 "Multivariate Analysis of Dichotomized Variables." American Journal of Sociology 70 (January).

Cronbach, Lee J.
1960 Essentials of Psychological Testing, 2nd ed., New York: Harper; 3rd ed., 1970.

Cronbach, L.J., and G.C. Gleser
1965 Psychological Tests and Personnel Decisions. Chicago: University of Illinois Press.

Deming, Edwards
1944 "On Errors in Surveys." American Sociological Review 9: 359-369. Reprinted as Ch. 31, pp. 341-358 in Billy J. Franklin and Harold W. Osborne, eds., Research Methods: Issues and Insights, Belmont, CA: Wadsworth, 1971.

Denzin, Norman K.
1970 The Research Act: A Theoretical Introduction to Sociological Methods. Chicago: Aldine.

Denzin, Norman K. (ed.)
1978 Sociological Methods: A Sourcebook. New York: McGraw Hill. Ch. 17, W. Edwards Deming, "On Errors in Surveys," 233-247.

Edwards, Allen L.
1957 Techniques of Attitude Scale Construction. New York: Appleton-Century-Crofts.

Ferguson, George A.
1959 Statistical Analysis in Psychology and Education. N.Y.: McGraw Hill.

Festinger, Leon, and Daniel Katz (eds.)
1953 Research Methods in the Behavioral Sciences. New York: Holt, Rinehart and Winston.

Fienberg, Stephen E.
1978 The Analysis of Cross-Classified Categorical Data. Cambridge, Mass.: The MIT Press.

Filstead, William J. (ed.)
1970 Qualitative Methodology. Chicago: Markham.

Finsterbusch, Kurt, and Annabelle Bender Motz
1980 Social Research for Policy Decisions. Belmont, CA: Wadsworth Publishing Co.

Glaser, Barney G., and Anselm L. Strauss
1967 The Discovery of Grounded Theory. Chicago: Aldine.

Goodman, Leo A.
1963 "On Methods for Comparing Contingency Tables." Journal of the Royal Statistical Society, Series A, 126: 94-108.

Goodman, Leo A.
1964 "Simple Methods for Analyzing Three-Factor Interaction in Contingency Tables." Journal of the American Statistical Association 59: 319-352.

Goodman, Leo A., and William H. Kruskal
1954 "Measures of Association for Cross-Classifications." Journal of the American Statistical Association 49: 732-764.

Hanushek, Eric A., and John E. Jackson
1977 Statistical Methods for Social Scientists. New York: Academic Press.

Hartwig, Frederick, and Brian E. Dearing
1979 Exploratory Data Analysis. Beverly Hills, CA: Sage.

Hays, William L.
1963 Statistics for Psychologists. New York: Holt, Rinehart and Winston.

Huff, Darrell
1964 How to Lie with Statistics. N.Y.: Norton.

Hyman, Herbert H.
1972 Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities. New York: Wiley.

Kahn, Robert L., and Charles F. Cannell
1957 The Dynamics of Interviewing. New York: Wiley. Ch. 6, The Design of Questionnaires. Plus 5 situational factors, p. 132.

Knoke, David, and Peter J. Burke
1980 Log-Linear Models. Beverly Hills, CA: Sage Publications. 80 p.

Lazarsfeld, Paul F., Ann K. Pasanella, and Morris Rosenberg (eds.)
1972 Continuities in the Language of Social Research. N.Y.: Free Press.

Lindner, William A.
1979 Statistics. Menlo Park, CA: Benjamin/Cummings Publishing Co., Inc.

McCall, George J., and J.L. Simmons (eds.)
1969 Issues in Participant Observation. Reading, MA: Addison-Wesley.

McNeil, Donald R.
1977 Interactive Data Analysis, A Practical Primer. New York: Wiley.

McNemer, Q.
1969 Psychological Statistics, 4th ed. New York: Wiley.

Miller, Delbert C.
1970 Handbook of Research Design and Social Measurement, 2nd ed. New York: David McKay.

Moiser, Charles I.
1947 "A Critical Examination of the Concepts of Face Validity." Educational and Psychological Measurement 7: 191-205.

Mosteller, Frederick, and John W. Tukey
1977 Data Analysis and Regression, A Second Course in Statistics. Reading, MA: Addison-Wesley.

Payne, Stanley L.
1951 The Art of Asking Questions. Princeton, N.J.: Princeton University Press. pp. 228-237, A concise checklist of 100 considerations.

Lofland, John
1971 Analyzing Social Settings: A Guide to Qualitative Observation and Analysis. Belmont, CA: Wadsworth Publishing Co.

Porter, Alan L., and Frederick A. Rossini
1980 "Technology Assessment/Environmental Impact Assessment: Toward Integrated Impact Assessment." IEEE Transactions on Systems, Man, and Cybernetics. Vol.SMC-10, No. 8 (August): 417-424.

Reynolds, H.T.
1977 The Analysis of Cross-Classifications. New York: Free Press.

Rosenberg, Morris
1968 The Logic of Survey Analysis. N.Y.: Basic Books.

Sieber, Sam D.
1973 "The Integration of Fieldwork and Survey Methods." American Journal of Sociology 78, 6 (May): 1335-1359.
Reprinted as reading 27 pp. 358-380 in Norman Denzin (ed.), Sociological Methods, A Sourcebook, 2nd ed., New York: McGraw Hill, 1978.

Siegel, Sidney
1956 Nonparametric Statistics for the Behavioral Sciences. N.Y.: McGraw Hill.

Tripodi, Tony, Phillip Fellin, and Henry J. Meyer
1969 The Asssessment of Social Research: Guidelines for the Use of Research in Social Work and Social Science. Itasca, IL: F.E. Peacock.

Tukey, John W.
1977 Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Weiss, Carol H. (ed.)
1972 Evaluating Action Programs: Readings in Social Action and Education.
Boston: Allyn and Bacon
Edward Suchman, "Action for What? A Critique of Evaluation Research," pp. 52-84.
Michael Scriven, "The Methodology of Evaluation." pp. 123-136.

Wonnacott, Thomas, and Ronald J. Wonnacott
1977 Introductory Statistics, Third ed., New York: John Wiley & Sons.

Younger, Mary Sue
1979 Handbook for Linear Regression. Belmont, CA: Duxbury Press (Wadsworth).

Zelditch, Morris, Jr.
1962 "Some Methodological Problems of Field Studies." American Journal of Sociology 67 (March): 566-576.
Reprinted as Ch. 32, pp. 495-511 in Norman Denzin (ed.), Sociological Methods, A Sourcebook, Chicago, Aldine, 1970.

Zuwailif, Fadil H.
1974 General Applied Statistics, second ed., Reading, MA: Addison-Wesley.

II. Alternate Teaching Approaches and Problems in Higher Education

Blackburn, Robert T., Glenn R. Pellino, Alice Boberg, and Colman O'Connell
1980 "Are Instructional Improvement Programs Off-Target?" Current Issues in Higher Education 2, 1 (September 3): 32-48. Issue title: Improving Teaching and Institutional Quality. Published by the American Association for Higher Education.

Becker, Howard, et al
n.a. Making the Grade. New York: Wiley.

Bloom, Benjamin S.,
1956 Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook I: Cognitive Domain. New York: David McKay. pp. 201-207.

Brubacher, John S.
1965 Bases for Policy in Higher Education. New York: McGraw-Hill.

Corwin, Ronald G.
1974 Education in Crisis: A Sociological Analysis of Schools and Universities in Transition. New York: Wiley.
ch. 7: Strategies for Educational Reform: 317-364.

Ernst, Ken
1972 Games Students Play. Millbrae, CA: Celestial Arts Publishing.

Kirschenbaum, Howard, Sidney B. Simon, and Rodney W. Napier
1971 Wad-ja-get?: The Grading Game in American Education. New York: Hart Publishing Company.

Mander, Jerry
1978 Four Arguments for the Elimination of Television. New York: Morrow.

Mauksch, Hans O.
1980 "What are the Obstacles to Improving Quality Teaching?" Current Issues in Higher Education 2, 1 (September 3): 49-56. Issue title: Improving Teaching and Institutional Quality. Published by the American Association for Higher Education.

McKeachie, Wilbert J.
1969 Teaching Tips. Lexington, Mass.: D.C. Heath.

Michaels, James W.
1978 "Effects of Innovative Course Procedures on Student Outcomes." Virginia Polytechnic Institute and State University. Xerox paper.

Miller, H.G., R.G. Williams, and T.M. Haladyna
1978 Beyond Facts: Objective Ways to Measure Thinking. Englewood Cliffs, N.J.: Educational Technology.

Miller, John P.
1976 Humanizing the Classroom. New York: Praeger.

Morris, William H. (ed.)
1970 Effective College Teaching: The Quest for Relevance. Washington, D.C.: American Association for Higher Education.

Novak, Michael
1972 The Rise of the Unmeltable Ethnics. New York: MacMillan.

Rogers, Carl R.
1969 Freedom to Learn. Columbus, OH: Charles E. Merrill.

Schwartz, Barry N. (ed.)
1972 Affirmative Education. Englewood Cliffs, N.J.: Prentice-Hall.

Scott, John Anthony
1972 Teaching for a Change. New York: Bantam Books.

Slawski, Carl
1972 "Matching of Learning and Teaching Styles: A Typology of Social Distance in Four Cultures." Pacific Sociological Association meetings, Portland, Oregon (April 13). ERIC document # ED 162 587, April 1979.

Unknown author
n.a. Applying Humanistic Techniques to Classroom Situations. ch. 5 in ? Columbus, OH: Charles E. Merrill.

A. Journals on Alternate Education and Research

Change
New Directions in Teaching. Bowling Green State U., Dept. of Ed.
Media Mix, Palatine, IL.
Media and methods. Philadelphia
Simulation/Gaming/News.

Memo to the Faculty. Ann Arbor: Center for research on learning and teaching.

Improving College and university teaching. Ore. State U.

Evaluation Comment. Center for the study of evaluation, 145 Moore Hall, 405 Hilgard Ave. L.A., CA 90024, free.

Teaching-Learning Issues. Learning Research Center, U. of Tennessee, Knoxville, free.

The Research Reporter. Center for Research and Development in Higher Education. UC Berkeley, free.

CIUE Notes. Center for Improvement of Undergraduate Education. Cornell U., free.

III. Student Evaluation of Instruction and its Validity

Astin, Alexander W.
1980 "Assessing Student Development: Improving Faculty Teaching Skills." American Association for Higher Education Bulletin 32, 6 (April): 13.

Aubrecht, Judith D.
1979 "Are Student Ratings of Teacher Effectiveness Valid?" Manhattan, KS: Kansas State University. Center for Faculty Evaluation and Development, Idea Paper No. 2 (November).

Aubrecht, Judith D.
1981 "Reliability, Validity and Generalizability of Student Ratings of Instruction." Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development. Idea Paper no. 6.

Armour, Robert
1979 "What Do They Expect of Me?" Point of View in The Chronicle of Higher Education (October 15): 48.

Barnoski, Robert P., and Alan L. Sockloff
1976 "A Validation Study of the Faculty and Course Evaluation (FACE) Instrument." Educational and Psychological Measurement 36, 2: 391-400.

Berliner, David C. (Far West Lab for ER&D)
1976 "Impediments to the Study of Teacher Effectiveness." Journal of Teacher Effectiveness." Journal of Teacher Education 27, 1 (Spring): 5-13.

Buros, Oscar Krisen (ed.)
1978 The Eighth Mental Measurements Yearbook, volume I, Highland Park, N.J.: The Gryphen Press.
Reviews in section 398 on Student Instructional Report, by Frank Costin, pp. 534-535; and by William C. McGaghie, pp. 535-536.

Bryson, Rebecca
1974 Teacher Evaluations and Student Learning: A Reexamination. Journal of Educational Research 68, 1 (September): 12-14.

Centra, John A.
1973 "The Student as Godfather? The Impact of Student Ratings on Academia." ETS Research Memorandum 73-8.

Centra, John A.
1976 "Two Studies on the Validity of the Student Instructional Report." I. "Student Ratings of Instruction and Their Relationship to Student Learning," by John A. Centra and Bert Rose. II. "The Relationship Between Student, Teacher, and Course Characteristics and Student Ratings of Teacher Effectiveness." SIR Report No. 4. Princeton, N.J.: Educational Testing Service.

Centra, John A.
1979 Determining Faculty Effectiveness. San Francisco: Jossey-Bass.

Chronicle of Higher Education
1979 Various Opinion articles and letters to the editor:
1981 November 13, 1979, p. 21: October 7 and November 11, 1981.

Cohen, Arthur M., and Florence Brawer
1969 Measuring Faculty Performance. Washington, D.C.: American Association of Junior Colleges.

Costin, Frank, William T. Greenough, and Robert J. Menges
1971 "Student Ratings of College Teaching: Reliability, Validity, and Usefulness." Review of Educational Research 41, 5 (December): 511-535.

Curry, Jerome
1976 "Students----generally lack the critical ability necessary for faculty evaluation." College Student Journal 10: 306-311.

Doyle, Kenneth O., Jr.
1975 Student Evaluation of Instruction. Lexington, MA: Lexington.

Doyle, Kenneth O., Jr., and S.E. Whitely
1974 "Student Ratings as Criteria for Effective Teaching." American Educational Research Journal 11 (Summer): 259-274.

Eble, Kenneth E., et al
1970 The Recognition and Evaluation of Teaching. Salt Lake City, Utah: Project to Improve College Teaching.

Educational Testing Service, Student Instructional Report
no. 1: "The Student Instructional Report: Its Development and Uses.
no. 3: "The Student Instructional Report
1. Comparisons with Alumni Ratings
2. Item Reliabilities
3. The Factor Stucture

French-Lazovik, Grace
c. 1975 "Evaluation of College Teaching: Guidelines for Summative and Formative Procedures." Washington, D.C.: Association of American Colleges, Occasional Paper. 12 pp.

Frey, Peter W.
1974 "The Ongoing Debate: Student Evaluation of Teaching." Change (February): 47-48 & 64. Original paper "Student instructional ratings and faculty performance," presented at annual meetings of AERA, New Orleans, Feb.-Mar. 1973.

Frey, Peter W.
1976 "Validity of Student Instructional Ratings: Does Timing Matter?" Journal of Higher Education 47, 3 (May/June): 327-336.

Gessner, P.K.
1973 "Evaluation of Instruction." Science 180: 566-569.

Glasman, Naftaly S.
1976 "Evaluation of Instructors in Higher Education: An Administrative Function." Journal of Higher Education 47, 3 (May/June): 309-327.

Greenwood, Gordon E., and Howard Ramagli, Jr.
1980 "Alternatives to Student Ratings of College Teaching." Journal of Higher Education 51 (November/December): 673-684.

Hansen, Barbara
1977 "Teacher evaluation is immoral." Community College Review 4, 3(Winter): 4-7.

Hildebrand, Milton, Robert C. Wilson, and Evelyn R. Dienst
1971 Evaluating University Teaching. Berkeley, CA: Center for Research and Development in Higher Education, University of California.

Hildebrand, Milton
1972 "How to Recommend Promotion for a Mediocre Teacher Without Actually Lying." Journal of Higher Education 43: 44-62.

Hodgkinson, Harold L.
1971 "Assessment and Reward Systems." chap. 4, pp. 47-54 in G. Kerry Smith, ed., New Teaching, New Learning.

Hogan, T.P.
1973 "Similarity of Student Ratings Across Instructors, Courses, and Time." Research in Higher Education 1: 149-154.

Kiernan, Irene R.
1975 "Student evaluations re-evaluated." Community and Junior College Journal 45, 7 (April): 25-27.

Kulik, James A., and Chen-Lin C. Kulik
1974 "Student Ratings of Instruction." Teaching of Psychology 1 (December): 51-57.

Kulik, James A., and Chen-Lin C. Kulik
1979 "College Teaching." In Penelope P. Peterson, and H.J. Walberg (eds), Research on Teaching. Berkeley: McCutchan, pp. 70-93.

Kulik, James A., and Wilbert J. McKeachie
1975 "The Evaluation of Teachers in Higher Education." In Fred N. Kerlinger (ed.), Review of Research in Education 3, Itasca, IL: Peacock.

Kulik, J.A., and W.J. McKeachie
1975 "The Evaluation of Teachers in Higher Education." Review of Research in Education 3: 210-240.

Landis, Larry M., and Ellen B. Pirro
1976 "Student Evaluation and the Rating of Course and Instructors: An Analysis of the Impact of Selected Student and Course Variables." Paper given at the Midwest Sociological Society, St. Louis (April).

Machlup, Fritz
1979 "Poor Learning from Good Teachers." Academe (October): 376-380.

Marsh, Herbert W.
1977 "The Validity of Students' Evaluations: Classroom Evaluations of Instructors Independently Nominated as Best and Worst Teachers by Graduating Seniors." American Educational Research Journal 14, 4 (Fall): 441-447.

Mastropaolo, Joseph
1972(?) "Subjective and Objective Evaluation of Faculty." Long Beach: California State University. Report in ditto form.

McKeachie, W.J.
1963 "Research on Teaching at the College and University Level." In N.L. Gage (ed.), Handbook of Research on Teaching. Chicago: Rand McNally, pp. 1118-1172.

McKeachie, Wilbert J.
1979 "Student Ratings of Faculty: A Reprise." Academe (October): 384-397.

McKeachie, Wilbert J., Yi-Guang Lin, and William Mann
1971 "Student Ratings of Teacher Effectiveness: Validity Studies." American Educational Research Journal 8, 3 (May): 435-445.

Menges, Robert J.
197_ "The New Reporters: Students Rate Instruction."

Miller, Richard I.
1972 Evaluating Faculty Performance. San Francisco: Jossey-Bass.

Millman, Jason (ed.)
1981 Handbook of Teacher Evaluation. Beverly Hills, CA: Sage.
articles
ch. 8, Student Ratings of Instruction, L.M. Aleamoni, pp. 110-145.
ch. 9, Student Achievement as a Measure of Teacher Competence, Jason Millman, 146-166.
ch. 15, Politics of Teacher Evaluation, John D. McNeil, UCLA, 272-291
ch. 16, The Political Realities of Teacher Evaluation, Mary Louise Armiger, 292-302.
ch. 17, Fairness and the Legal Context of Teacher Evaluation, Kenneth Strike and Barry Bull, 303-343.

Powell, Robert
1978 "Faculty Rating Scale Validity: The Selling of a Myth." College English 39, 5 (January): 616-629.

Raskin, Betty Lou, and Patricia R. Plante
1979 "The Student Devaluation of Teachers." Academe (October): 381-383.

Peterson, Penelope L., and Herbert J. Walberg (eds.)
1979 Research on Teaching: Concepts, Findings, and Implications. Berkeley, CA: McCutchan Publishing Corporation.

Renner, Richard R.
1981 "Comparing Professors: How Student Ratings Contribute to the Decline in Quality of Higher Education." Phi Delta Kappan (October): 128-130.

Rodin, M., and B. Rodin
1972 "Student Evaluations of Teachers." Science 177: 1164-1166.

Sabine, Gordon A.
1971 "How Students Rate Their Schools and Teachers." Washington, D.C.: The National Association of Secondary School Principals.

Schaffer, Ronald
1976 "Quantitative Student Teacher Evaluations." Editorial in UPC Advocate 6, 6 (March): 7.

Scheck, Dennis C.
1978 "The Use and Abuse of Student Evaluations of Teaching Effectiveness in Higher Education." College Student Journal Monograph 12, 3, part 2 (Fall): 13p.

Scheck, Dennis C.
1979 "Peer Evaluations of College Instructors' Teaching Effectiveness: Some Critical Concerns." College Student Journal 13, 2 (Summer): 122-125.

Schuh, Allen J., and Michael A. Criveili
1973 "An Animadversion error in student evaluations of faculty teaching effectiveness." Journal of Applied Psychology 58: 259-260.

Scriven, Michael
1980 "Summative Teacher Evaluation." San Francisco, The Evaluation Institute. Xerox copy of version 7 from author.

Sheehan, Daniel S.
1975 "On the Invalidity of Student Ratings for Administrative Personnel Decisions." Journal of Higher Education 46, 6 (November/December): 687-700.

Shingles, Richard D.
1977 "Faculty Ratings: Procedures for Interpreting Student Evaluations." American Educational Research Journal 14, 4 (Fall): 459-470.

Shipka, Thomas A.
1977 "Is Peer Review Viable Today?" pp. 41-45 in Aaron Levenstion (ed.), Collective Bargaining and the Future of Higher Education, Proceedings of the Fifth Annual Conference, (April): City University of New York, Baruch College, National Center for the Study of Collective Bargaining in Higher Education.

Shore, Jane E.
1978 "Faculty Evaluation in Higher Education." The Generator (of the American Educational Research Association) 8, 3 (Fall): 1-7.

Slawski, Carl
1976 "Dilemmas in Evaluating the Learning Relationship." Midwest Sociological Society St. Louis, MO: (April). ERIC # SP 010 881, ED #137 244, 1977.

Slawski, Carl
1979 "Selecting Student Jurors for Evaluating Teachers." Midwest Sociological Society, Minneapolis (April). ERIC #ED 157 452.

Sullivan, Arthur M., and Graham R. Skanes
1974 "Validity of Student Evaluation of Teaching and the Characteristics of Successful Instructors." Journal of Educational Psychology 66, 4: 584-590.

Thayer, Frederick C.
1979 "Student evaluations of professors: negative and positive impacts." Letter to the editor of The Chronicle of Higher Education, November 13, 1979 (in response to Robert Armour's Point of View (Oct. 15)) p. 21.

Thayer, Robert E., et al
1972 "Final Report and Recommendations of Ad Hoc Committee on Student Evaluation of Faculty." Presented to the California State University, Long Beach, Academic Senate (c. March). 20 pp.

Thorne, Gaylord L.
1980 "Student Ratings of Instructors: From Scores to Administrative Decisions." Journal of Higher Education 51, 2: 207-214.

Trent, James W., and Arthur M. Cohen
1973 "Research on Teaching in Higher Education." Chap. 32 in R.M.W. Travers (ed.), Second Handbook of Research on Teaching. Chicago: Rand McNally, pp. 997-1071. Subsection on "Evaluation of Teaching" by Arthur M. Cohen, James W. Trent, and Clare Rose, pp. 1041-1057.

Turner, Richard L., and Robert P. Thompson
1974 "Relationships Between College Student Ratings of Instructors and Residual Learning." Presented at AERA, Chicago: April 1974. ERIC document #ED 090 826.

Whitley, Susan E., and Kenneth O. Doyle
1976 "Implicit Theories in Student Ratings." American Fducational Research Journal 13, 4 (Fall): 241-253.

Williams, Reed G., and John E. Ware, Jr.
1977 "An Extended Visit with Dr. Fox: Validity of Student Satisfaction with Instruction Ratings after Repeated Exposures to a Lecturer," American Educational Research Journal 14, 4 (Fall): 449-457.

Wittrock, M.C., and Arthur A. Lumsdaine
1977 "Instructional Psychology." Annual Review of Psychology 28: 417-459.

IV. Legal Aspects

American Association of University Professors
1977 "Policy Documents and Reports." Washington, D.C.: AAUP.

Anonymous
n.d. Employment Practices Guide. Chicago: Commerce Clearing House.

Baird, J., and M.R. McArthur
1976 "Constitutional Due Process and the Negotiation of Grievance Procedures in Public Employment." Journal of Law and Education 209, 5(?).

California State University, Long Beach
1981-82 Faculty Handbook. Long Beach: CSULB.

Code of Federal Regulations
1981 Vol. 29, parts 900 to 1899, July 1, 1981. Labor.

Department of Water and Power, City of Los Angeles
1974 Civil Service Rules of the City of Los Angeles, Employee Handbook.

Fair Employment Practice Manual
serial Labor Relations Reporter. Washington, D.C.: Bureau of National Affairs.

Federal Register
1978 "Uniform Guidelines on Employee Selection Procedures." Federal Register (August 25) pp. 38296-38297.

Gray, Mary W., and Alice T. Schafer
1981 "Guidelines for Equality: A Proposal." Academe 67, 6 (December): 351-354.

Kaplin, William A.
1978 The Law of Higher Education: Legal Implications of Administrative Decision Making. San Francisco: Jossey-Bass.

Kaplin, William A.
1980 The Law of Higher Education 1980. San Francisco: Jossey-Bass.

Munsee vs. Horn
1977 West's California Reporter 139, Supreme Court, Courts of Appeal, 66 Cal. App. 3d, pp. 373-377, (June 30).

Nairn, Allan, and associates
1980 The Reign of ETS: The Corporation That Makes Up Minds. Washington, D.C.: Learning Research Project.

Peterson, Donovan, and Annie Ward
1980 Due Process in Teacher Evaluation. Washington, D.C.: University Press of America, Inc.

Peterson, L.J., R.A. Rossmiller, and M.M. Volz
1978 The Law and Public School Operation. New York: Harper & Row.

Polishook, Irwin H.
1977 "Peer Judgment and Due Process." Pp. 36-40 in Collective Bargaining and the Future of Higher Education by Aaron Levenstein (ed.), Proceedings of the Fifth Annual Conference, (April): City University of New York, Baruch College, National Center for the Study of Collective Bargaining in Higher Education.

Psychological Corporation
1978 Summaries of Court Decisions on Employment & Testing. New York City: Author.

Siegel, Jerome
1980 Personnel Testing Under EEO. New York: AMACOM.

United States Equal Employment Opportunity Commission
1974 Affirmative Action and Equal Employment, A Guidebook for Employers, 2 vols., Washington, D.C.: Government Printing Office.

United States Office of Personnel Management
1980 "A Guide for Improving Performance Appraisal." Washington, D.C.: Government Printing Office.

United States Office of Personnel Management
1981 Manager's Handbook. Washington, D.C.: Government Printing Office.